200 ans après Laplace, l'essor des méthodes bayésiennes d'analyse des données

D. Boilley, Y. Lallouet

200 ans après Laplace, l'essor des méthodes bayésiennes d'analyse des données

par **David BOILLEY**
Normandie Université - 14000 Caen
GANIL - 14000 Caen
boilley@ganil.fr
et **Yoann LALLOUET**
Lycée Malherbe - 14000 Caen
yoann.lallouet@ac-caen.fr

Résumé

Initiée par Pierre-Simon de Laplace il y a deux cents ans, l'analyse bayésienne des données expérimentales est en plein essor. Elle a fait son entrée dans certaines normes internationales de métrologie. Le but de cet article est d'introduire simplement le changement philosophique sous-jacent. Des applications à différents types de mesurages sont présentées.

1. Incertitude sur la masse de Saturne

Dans la troisième édition de sa *Théorie analytique des probabilités*, Pierre-Simon de Laplace, féru de mécanique céleste, souhaite mettre en avant la précision des observations et des théories : *« Mais il restait à déterminer la probabilité des erreurs que cette correction laisse encore à craindre : c'est ce que la méthode que je viens d'exposer fait connaître. Pour en donner quelques applications intéressantes, j'ai profité de l'immense travail que M. Bouvard vient de terminer sur les mouvemens de Jupiter et de Saturne, dont il a construit des tables très précises. [...] Ses calculs lui donnent la masse de Saturne égale à la 3512e partie de celle du Soleil. En leur appliquant mes formules de probabilité, je trouve qu'il y a onze mille à parier contre un, que l'erreur de ce résultat n'est pas un centième de sa valeur, ou, ce qui revient à très peu près au même, qu'après un siècle de nouvelles observations ajoutées aux précédentes, et discutées de la même manière, le nouveau résultat ne différera pas d'un centième de celui de M. Bouvard. »*

En langage moderne, la probabilité que l'incertitude relative sur la masse de Saturne soit supérieure à 1% est de 1/11 000. La masse de Saturne donnée par la NASA en 2004 est à moins de un centième de celle de M. Bouvard. Laplace a gagné son pari !

Le calcul est détaillé dans son *Mémoire sur l'application du calcul des probabilités à la philosophie naturelle*, daté de 1815, il y a tout juste 200 ans. Qu'entend-il par *« mes formules de probabilités »* ? Ce que l'on appelle, de nos jours, la formule de Bayes, du nom du Révérend Bayes, publiée à titre posthume par son ami Richard Price en 1763, il y a un peu plus de 250 ans (l'article original est d'une lecture difficile, mais Stigler explique la démarche en langage mathématique moderne). Laplace l'a retrouvée indépendamment, en 1774, de manière beaucoup plus générale et l'a publiée dans son *Mémoire sur la probabilité des causes par les évènemens*. Il n'avait que 25 ans. Il est le premier à l'appliquer à l'étude des données expérimentales.

Nous allons, ici, non pas reproduire les calculs de Laplace, mais présenter de manière ludique les concepts mathématiques utilisés et proposer quelques applications simples en métrologie.

Les textes de Laplace sont disponibles en ligne sur de nombreux sites, dont Gallica

de la BNF.

2. Quelques rappels sur les probabilités

Daniel Kahneman, psychologue américain, prix Nobel d'économie, propose le quizz suivant :

Linda est une employée de banque âgée de 31 ans, célibataire, franche et intelligente. Elle a fait des études de philosophie. Quand elle était étudiante, elle se sentait concernée par les problèmes de discrimination et d'injustice sociale. Elle participait aussi à des manifestations antinucléaires. Qu'est-ce qui est le plus probable :

- *Linda est une employée de banque ;*
- *Linda est une employée de banque engagée dans le mouvement féministe ?*

C'est, bien évidemment, la première réponse qui est moins restrictive que la seconde, même si le texte nous incite à choisir la deuxième. Dans son ouvrage passionnant sur le fonctionnement du cerveau et les biais cognitifs qu'il induit, Kahneman met en avant les deux systèmes de la pensée : le premier, rapide et intuitif, peut nous induire en erreur. Le second, plus lent et réfléchi, nous permet de faire des mathématiques.

Si, pour Laplace, *« la théorie des probabilités n'est que le bon sens réduit au calcul »,* c'est au deuxième système qu'il fait implicitement mention et auquel nous ferons appel.

Prenons un jeu de 52 cartes pour se rappeler les règles concernant les probabilités. La probabilité de tirer un as de cœur est égale à la probabilité de tirer un as multipliée par la probabilité de tirer un cœur. Mathématiquement, cela s'écrit

$$\mathrm{P(A\ et\ C)} = \mathrm{P(A)}\times\mathrm{P(C)} = \frac{1}{13}\times\frac{1}{4} = \frac{1}{52}.$$

Ici, « as » et « cœur » sont indépendants. Cette formule est fausse si on l'applique au calcul de la probabilité de tirer un cœur rouge, car tous les cœurs sont rouges. « Cœur » et « rouge » ne sont pas indépendants. Il faut alors avoir recours aux probabilités conditionnelles. Mathématiquement, on écrit

$$\mathrm{P(C\ et\ R|H)} = \mathrm{P(C|H)}\times\mathrm{P(R|CH)}.$$

Cela se lit comme la probabilité d'avoir une carte qui est « cœur » et « rouge » est égale à la probabilité que la carte soit un cœur multipliée par la probabilité que la carte soit rouge si elle est un cœur. La condition « si elle est un cœur » est mise à droite de la barre verticale |. Dans l'expression mathématique ci-dessus, on a ajouté les hypothèses sous jacentes, notées *H*, qui pourraient être ici « si le jeu n'est pas truqué ».

Pour Laplace, « *quand deux évènemens dépendent l'un de l'autre, la probabilité de l'évènement composé est le produit de la probabilité du premier évènement, par la probabilité que cet évènement étant arrivé, l'autre arrivera. »* De nos jours, les lycéens connaissent cette loi au travers d'arbres de probabilité.

Mais attention, $P(C|RH) \neq P(R|CH)$. De même, la probabilité qu'il pleuve quand il y a des nuages n'est pas égale à la probabilité qu'il y ait des nuages quand il pleut.

Pour accéder à la probabilité inverse, on aurait pu aussi dire que la probabilité de

tirer une carte qui est « cœur » et « rouge » est égale à la probabilité que la carte soit rouge multipliée par la probabilité qu'elle soit un cœur si elle est rouge :

$$\mathrm{P(C\ et\ R|H)} = \mathrm{P(R|H)}\times\mathrm{P(C|RH)}.$$

Les deux équations précédentes, qui sont égales entre elles, conduisent à la première version du théorème de Bayes,

$$\mathrm{P(C|RH)} = \frac{\mathrm{P(C|H)}\times\mathrm{P(R|CH)}}{\mathrm{P(R|H)}},$$

qui permet de relier deux probabilités inverses. Nous présenterons plus loin l'intérêt de ce résultat. Pour l'appliquer à des cas concrets, on a encore besoin d'une autre relation mathématique liée à l'opération *ou*.

La probabilité de tirer un as dans un jeu de cartes est égale à la somme des probabilités de tirer un as dans chaque couleur. Cette propriété est vraie car les sous-ensembles sont disjoints, un as ne pouvant pas être à la fois de cœur et de carreau, et complets, les quatre couleurs couvrant toutes les possibilités. Evidemment, l'intérêt est limité avec cet exemple car les probabilités sont identiques pour chaque couleur. Prenons un autre exemple où la séparation est nécessaire.

Supposons qu'un piège peut attraper aussi bien des souris que des musaraignes, sans distinction sachant qu'il y a 60% de musaraignes. S'il y a 57% de femelles chez les souris et 53% chez les musaraignes, combien de chances a-t-on de capturer une femelle ? Comme un animal ne peut être à la fois souris et musaraigne, on a immédiatement, les initiales parlant d'elles-mêmes :

$$\mathrm{P(F)} = \mathrm{P(F\ et\ M)} + \mathrm{P(F\ et\ S)}.$$

En appliquant les formules précédentes, il vient

$$\mathrm{P(F)} = \mathrm{P(M)}\times\mathrm{P(F|M)} + \mathrm{P(S)}\times\mathrm{P(F|S)} = 0{,}6\times0{,}53 + 0{,}4\times0{,}57 = 54{,}6\%.$$

La Figure 1 présente ce résultat sous forme d'arbres des possibilités tel qu'ils sont enseignés au lycée. Cette dernière formule nous servira à calculer le dénominateur dans le théorème de Bayes.

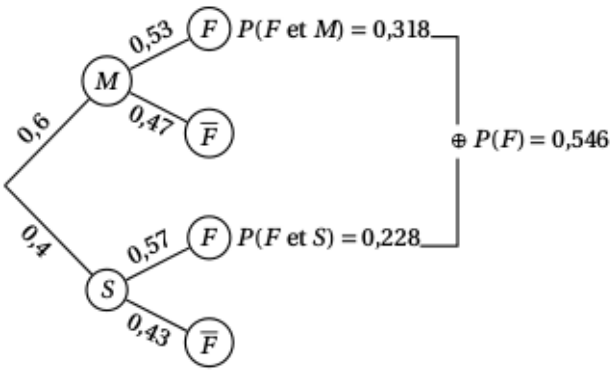

Figure 1 : *Arbre pondéré des possibilités permettant de trouver la probabilité que le piège à souris et musaraignes attrape une femelle. Les données sont dans le texte.*

Finalement, Laplace, dans son Mémoire sur la probabilité des causes par les évènements de 1774, présente le même théorème sous cette forme : *« Si un évènement peut être produit par un nombre n de causes différentes, les probabilités de l'existence de ces causes prises de l'évènement sont entre elles comme les*

probabilités de l'événement prises de ces causes, et la probabilité de l'existence de chacune d'elles est égale à la probabilité de l'évènement prise de cette cause, divisée par la somme de toutes les probabilités de l'évènement prises de chacune de ces causes. »

Ecrit en langage mathématique, c'est plus clair :

$$P(C|E) = \frac{P(E|C) \times P_{prior}(C)}{\sum P(E|C') \times P_{prior}(C')},$$

où E signifie « événement » et C, « cause ». On reviendra sur l'explication du théorème et des différents termes dans les exemples qui suivent. Ce théorème est au programme des classes préparatoires scientifiques.

3. La probabilité vue comme l'information

Les règles de base sur les calculs de probabilités ont été présentées ici sur des exemples où elles correspondent à des fréquences statistiques qui peuvent être estimées en répétant un grand nombre de fois l'expérience. Mais Cox a montré que les mêmes règles s'appliquaient à toutes les probabilités, même s'il n'y a pas une interprétation statistique sous-jacente. C'est le cas avec la plausibilité d'événements à laquelle on attribue une probabilité, comme, par exemple, la probabilité qu'il y ait de la vie sur Mars. Comme le note Laplace, la probabilité est relative en partie à notre ignorance, en partie à nos connaissances.

Faisons un autre jeu proposé dans le livre pédagogique de Sivia pour bien comprendre. Supposons que l'on a mis 7 boules noires dans une urne et 3 blanches. On mélange le tout et l'on tire une boule au hasard. Tout le monde s'accorde pour dire que la probabilité que la boule soit blanche est de 30% et qu'elle soit noire, de 70%. On cache cette boule sans l'avoir regardée et on en tire une deuxième. Connaissant la couleur de la deuxième boule, est-ce que cela change la probabilité issue du premier tirage ? Réfléchissez.

Un doute ? Supposez maintenant qu'il n'y avait que deux boules dans l'urne, une noire et une blanche. En déterminant la couleur de la deuxième, on connaît immédiatement la couleur de la première. Ainsi, dans l'exemple précédent, si la deuxième boule est blanche, la probabilité que la première soit blanche n'est plus que de 2/9. Celle d'être noire, de 7/9.

Le premier tirage n'est pas affecté par le deuxième tirage. La probabilité de tirer une boule blanche est bien de 30%. En revanche, l'information que l'on a sur le résultat de ce premier tirage est affectée par le tirage suivant. Les probabilités de 2/9 et 7/9 correspondent à la plausibilité, pas à la fréquence.

En métrologie, l'étude des incertitudes préconisée par le *Guide pour l'expression de l'incertitude de mesure* (GUM) repose sur les deux interprétations. Les incertitudes de type A, issues d'une étude statistique, relèvent de l'approche dite *fréquentiste*, alors que les incertitudes de type B relèvent de l'approche dite *plausibiliste*. En effet, pour déterminer une incertitude de lecture, par exemple, on ne procède pas à une répétition des lectures suivie par une étude statistique. On estime par d'autres méthodes que la distribution de probabilité doit être, par exemple, rectangulaire ou triangulaire et l'on en déduit l'écart-type. Voir notre article sur l'estimation de l'incertitude.

4. Un exemple simple d'inférence bayésienne

Prenons encore un exemple pour voir comment on utilise le théorème de Bayes pour en déduire une information recherchée en prenant en compte, à la fois le résultat d'un mesurage et les autres connaissances.

Avant de proposer un traitement lourd à un malade, un médecin va faire procéder à un examen médical qui n'est jamais complètement fiable. Supposons que 90% des malades testés ont un résultat positif et qu'il y a 5% de faux positifs, c'est à dire des personnes saines ayant un résultat positif au test. Sachant que le test, qui correspond à un mesurage unique, est positif, quelle est la probabilité que le patient ait la maladie recherchée ?

On peut difficilement répéter les mesurages dans ce cas là. On connaît la probabilité que le patient ait un test positif (que l'on notera *P*) s'il est malade (que l'on notera *M*) et l'on veut connaître la probabilité inverse, à savoir la probabilité d'être malade si le résultat est positif. C'est le théorème de Bayes qui va permettre d'inverser les probabilités :

$$P(M|P) = \frac{P_{prior}(M) \times P(P|M)}{P(P)}.$$

Dans cette équation, $P(P|M)$ est la probabilité que le test soit positif si le patient a la maladie. Elle est généralement déterminée par le fabricant du test. $P_{prior}(M)$ est appelé le prior et correspond à la probabilité estimée par le médecin que le patient soit malade avant que le test soit effectué. Enfin, $P(P)$, au dénominateur, correspond à toutes les façons d'avoir un test positif :

$$P(P) = P_{prior}(M) \times P(P|M) + P_{prior}(\overline{M}) \times P(P|\overline{M}).$$

Ici, $\overline{M}$ signifie non-*M*, c'est à dire que le patient n'a pas la maladie recherchée. $P(P)$ correspond à la normalisation. La Figure 2 reprend cet exemple d'inférence bayésienne avec un arbre des possibilités connu des lycéens.

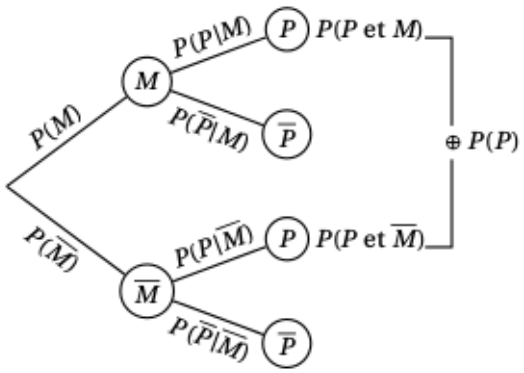

Figure 2 : *Arbre de possibilités permettant de déterminer la probabilité qu'un malade ait la maladie M recherchée en cas de test positif* $P(M|P)$ *connaissant la probabilité que le test soit positif si le patient était malade* $P(P|M)$ *et s'il ne l'était pas* $P(P|\overline{M})$.

Le médecin n'est pas très sûr de lui et fixe donc le prior, à savoir la probabilité issue de son diagnostic que le patient ait la maladie recherchée, à $P_{prior}(M) = 0{,}5$. L'application de la formule de Bayes, en cas de test positif, conduit à $P(M|P) = 0{,}95$.

Si le patient a tous les symptômes de la maladie en question et que le médecin fixe

le prior à $P_{prior}(M) = 0{,}8$, on aura alors, après le test, $P(M|P) = 0{,}99$. Il y a peu de chances que le médecin et le test se trompent simultanément. D'où le résultat.

Imaginons maintenant une situation où il y a contradiction. C'est le cas, par exemple, si le test est proposé dans le cadre d'un dépistage et que le patient ne présente aucun des symptômes attendus. On peut alors prendre comme prior la fréquence d'occurrence de cette maladie parmi la population. Supposons qu'il s'agissent d'une maladie assez rare et que $P_{prior}(M) = 0{,}01$. On obtient alors $P(M|P) = 0{,}15$. Un test positif ne signifie donc pas maladie dans ce cas. Il y aura peu de résultats positifs, mais parmi eux, 85% sont de faux positifs. Il ne faut donc pas s'affoler et refaire un test.

5. L'inférence bayésienne vue comme un apprentissage

L'approche bayésienne manquait aux Shadoks qui avaient calculé qu'ils avaient une chance sur un million que leur fusée leur permette d'échapper à leur planète. Ils se dépêchaient donc d'avoir 999 999 échecs pour avoir enfin un succès. Leur devise était alors : *« En essayant continuellement, on finit par réussir. Donc : plus ça rate, plus on a de chances que ça marche »*. Ils n'avaient pas appris à apprendre de leurs échecs…

On retire de l'exemple d'inférence que la formule de Bayes est une formule d'apprentissage : elle permet d'inclure la connaissance apportée par le résultat du test dans la probabilité qui prend en compte notre ignorance et nos connaissances. Les compagnies d'assurance l'ont compris depuis longtemps et utilisent le formalisme bayésien pour actualiser la probabilité d'occurrence de risques et fixer leurs tarifs. Alan Turing et ses collaborateurs ont aussi utilisé les inférences bayésiennes pour décrypter le code Enigma pendant la seconde guerre mondiale. Stanislas Dahaene, spécialiste des sciences cognitives, explique, dans ses cours au collège de France, que les processus d'apprentissage des bébés sont bayésiens.

Le mesurage peut être vu comme un processus d'apprentissage qui permet d'obtenir une information sur la grandeur mesurée. En 1883, William Thomson, alias Lord Kelvin, écrivait dans le premier volume de son recueil de cours et conférences : *« Quand on peut mesurer ce dont on parle, et l'exprimer avec des nombres, on sait quelque chose ; mais quand on ne peut pas l'exprimer avec des nombres, la connaissance est maigre et peu satisfaisante ; cela peut être le début de la connaissance, mais vous n'aurez, dans votre esprit, à peine progressé vers les sciences, et ce pour toutes les matières »* (cité par Stigler).

Il existe, cependant, de nombreuses situations en sciences où l'on ne peut pas multiplier les mesurages et l'on doit se contenter de l'information apportée par un seul mesurage. Comme on l'a vu, le résultat obtenu in fine est très sensible au choix du prior. C'est normal puisqu'il n'y a eu qu'un seul mesurage. Si l'on multiplie les tests, en prenant le posterior précédent comme prior du test suivant, le résultat final est de moins en moins sensible au choix du tout premier prior. On peut faire des tests mathématiques simples qui permettent de le vérifier. Voir, par exemple, le livre de Sivia.

Dans l'exemple d'inférence précédent, le prior issu du diagnostic du médecin semble très subjectif. Deux médecins n'auraient sûrement pas choisi la même probabilité. Les inférences bayésiennes sont parues trop subjectives pour être scientifiques et l'approche a eu ses ennemis. Il existe maintenant des règles claires.

6. Détermination du prior

Pour une information donnée, chaque personne doit proposer le même prior. Pour cela, il faut des règles.

Recommençons à jouer, à pile ou face cette fois-ci. Si l'on veut vérifier si la pièce est biaisée ou non, on n'a pas d'autre choix, au tout début, de supposer qu'elle ne l'est pas et d'affecter le même poids à chaque possibilité. Rappelons que les probabilités représentent l'état de nos connaissances et qu'elles ne mesurent pas forcément la fréquence d'un évènement, mais plutôt sa plausibilité. Nous avions fait la même hypothèse sous jacente dans les autres exemples ludiques précédents. Chaque carte avait la même probabilité d'être tirée.

Si l'on n'a pas de raison de penser que *A* est plus probable que *B*, alors, on écrit $P(A) = P(B)$. C'est le *principe de la raison insuffisante* de Jacob Bernoulli. En 1921, Keynes proposa une autre formulation : *principe d'indifférence*.

Ce principe est facilement applicable aux exemples ludiques, mais pas toujours aux cas physiques qui nous intéressent, comme Bernoulli l'avait déjà expliqué. Il considérait même que les jeux de hasards inventés par l'homme se devaient de satisfaire cette règle simple.

La méthode utilisée maintenant pour déterminer le prior repose sur la maximisation de l'entropie de l'information. L'application de ce principe fait aussi appel à des méthodes de calcul assez sophistiquées qui dépassent l'ambition de cet article. L'article de Jaynes est particulièrement lumineux sur le sujet. Dans les années 1980, Shore et Johnson ont démontré mathématiquement, à partir d'une axiomatique, qu'il s'agissait de la seule méthode correcte dans le cadre d'une inférence déductive.

Jérôme Segal dans son ouvrage sur l'histoire de la science de l'information a montré son lien ancien avec l'entropie physique. L'interprétation de ce lien est complexe et a évolué au cours de l'histoire.

Si la seule information disponible est que la valeur attendue est dans un intervalle donné, le principe de la raison insuffisante ou la maximisation de l'entropie conduisent à une distribution rectangulaire, avec une probabilité uniforme dans l'intervalle et nulle en dehors.

Et si la seule information disponible est la valeur moyenne et l'écart-type, la maximisation de l'entropie en prenant en compte ces contraintes (à l'aide de multiplicateurs de Lagrange) conduit à une distribution de probabilité gaussienne. Le calcul est fait dans les articles originaux de Shannon.

7. Application à la mesure de la radioactivité

Quand on souhaite connaître l'activité d'une source radioactive, on ne fait généralement qu'un seul mesurage qui permet d'en déduire la valeur recherchée et son incertitude. Comment s'y prend-on ?

On a des informations solides sur les propriétés physiques de la radioactivité, obtenues par de nombreux mesurages précédents. On sait, ainsi, que *N*, le nombre de coups détectés, satisfait à la distribution de Poisson,

$$P(N|\mu) = \frac{\mu^N}{N!} e^{-\mu}.$$

Les propriétés de cette loi sont bien connues : le nombre moyen de coups $< N >$ et

sa variance sont tous les deux égaux à μ.

Mais, lors d'un mesurage, on obtient *N* et l'on veut connaître μ. Et donc, connaissant un résultat de mesure N_1, fixé, on cherche $P(\mu|N_1)$. C'est donc l'inverse de la fonction de Poisson que l'on peut déterminer à l'aide du théorème de Bayes.

Avant de présenter le calcul, il est important de préciser qu'avec cette approche, les résultats de mesure ne fluctuent plus comme c'était le cas lors que l'on fait une analyse d'incertitude classique. Ils sont fixes et ce sont les grandeurs recherchées qui fluctuent. Ici, $P(\mu|N_1)$ représente la plausibilité d'obtenir une valeur donnée pour μ connaissant le résultat d'un mesurage N_1.

Le théorème de Bayes donne

$$P(\mu|N_1) = \frac{P_{prior}(\mu)\times P(N_1|\mu)}{P(N_1)}.$$

Si l'on n'a aucune information sur la valeur de μ, la seule chose que l'on sait avec certitude est que $\mu \geq 0$. On choisit alors comme prior une fonction constante entre 0 et l'infini car on n'a aucune raison de privilégier une valeur plutôt qu'une autre. Si l'échantillon analysé avait déjà fait l'objet d'un mesurage, on aurait pu prendre comme prior la fonction de densité de probabilité obtenue précédemment.

Le dénominateur, qui correspond à la normalisation, somme toutes les façons d'obtenir un résultat égal à N_1 :

$$P(N_1) = \int_0^{+\infty} P_{prior}(\mu)\times P(N_1|\mu)\, d\mu.$$

En intégrant N_1 fois par partie, on obtient finalement que

$$P(\mu|N_1) = \frac{\mu^{N_1}}{N_1!}e^{-\mu}.$$

On peut aussi aisément calculer la valeur moyenne et la variance de cette fonction et l'on trouve $N_1 + 1$ pour les deux. Cela justifie a posteriori que $P(N|\mu)$ et $P(\mu|N_1)$ diffèrent. Ces deux distributions n'ont pas la même moyenne, ni la même variance. En revanche, elles ont la même allure.

Cette approche est à la base de la nouvelle norme internationale ISO 11929 relative au mesurage de la radioactivité qui date de 2009 (2010 pour la version française).

8. Près d'un seuil

Supposons que l'on ait mesuré une grandeur et obtenu un résultat peu précis, avec, par exemple, une incertitude relative de 50% : $x = 2 \pm 1$, peu importe l'unité. S'il s'agit d'une grandeur que l'on sait positive, comme une masse, une concentration, une longueur… ce résultat pose un problème. En supposant que la distribution des valeurs est une gaussienne, environ 2,5% d'entre elles sont négatives, ce qui est impossible. Comment prendre en compte ce fait, qui est plus certain que le résultat de ce mesurage de piètre qualité ?

Imaginons qu'il s'agisse d'une masse. En fonction de la technique mise en œuvre, ce que l'on observe, c'est le déplacement d'une aiguille liée au plateau de la balance ou la déviation d'un faisceau dans le cas de la spectrométrie de masse, etc… Evidemment, les mesurages de masse conduisent généralement à des résultats plus

précis, même si, dans le cas de neutrinos ou de corps célestes, l'incertitude demeure très grande.

Notons x la grandeur directement mesurée et m la masse. Un modèle permet de connaître la déviation x en fonction de masses qui ont servi à l'étalonnage. Cela se traduit généralement pas une fonction mathématique, $x = f(m)$. Après le mesurage, on veut l'inverse, à savoir la masse m en fonction de la déviation x avec les incertitudes associées. C'est encore le théorème de Bayes qui va nous permettre de répondre :

$$P(m|x) = \frac{P_{prior}(m) \times P(x|m)}{\int P_{prior}(m) \times P(x|m) dm}.$$

Comme on sait que la masse est positive, on choisit comme prior la fonction de Heavyside $P_{prior}(m) = \theta(m)$ qui vaut 1 si $m > 0$ et 0 si $m < 0$.

Pour faire le calcul, il faut connaître la fonction de distribution des données expérimentales $P(x|m)$. Par soucis de simplicité, on va supposer qu'elle est gaussienne. Cela permet de faire un calcul analytique. Ainsi,

$$P(m|x) = \frac{P_{prior}(m) \times \exp\left(-\frac{(x - f(m))^2}{2u_x^2}\right)}{\int P_{prior}(m) \times \exp\left(-\frac{(x - f(m))^2}{2u_x^2}\right) dm}.$$

Enfin, si l'on suppose que la déviation est bien étalonnée en kilogramme ou toute autre unité de masse, et que l'on a simplement $m = x$, on peut calculer simplement $P(m|x)$ et en déduire $\overline{m}$ la valeur moyenne de la masse et l'incertitude associée. Les expressions mathématiques sont lourdes, nous ne les écrirons pas ici.

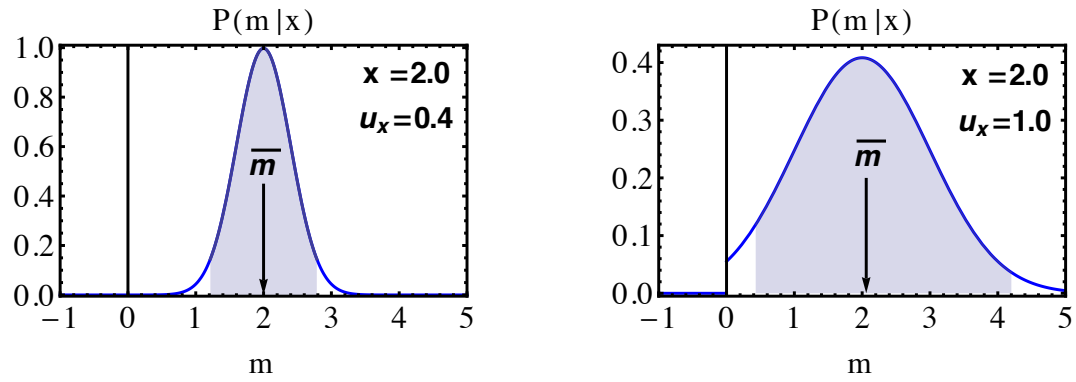

Figure 3 : *Représentation de la fonction densité de probabilité de la masse m pour une déviation x donnée et deux valeurs de l'incertitude associée. L'intervalle de confiance qui inclut 95% des valeurs est indiqué en grisé.*

Nous avons représenté, Figure 3, la fonction densité de probabilité $P(m|x)$ pour deux incertitudes différentes pour x. La valeur moyenne obtenue est indiquée par une flèche. La zone grisée correspond à l'intervalle qui inclut 95% des valeurs de m.

Si le résultat du mesurage a une faible incertitude relative, alors $\overline{m} = x$ et $P(m|x) \approx P(x)$. Cela ne remet pas en cause tout ce que l'on a appris. En revanche, si l'incertitude relative est grande, il faut corriger le résultat expérimental en prenant en compte le fait que la masse ne peut pas être négative. La valeur obtenue pour $\overline{m}$ n'est plus égale à x.

Ainsi, si $x = 2 \pm 1$, $\overline{m} = 2{,}055$. Il y a un léger décalage. La valeur moyenne n'est pas la plus probable et l'intervalle de confiance n'est plus symétrique par rapport à $\overline{m}$. Ici, l'intervalle qui inclut 95% des valeurs est $[0{,}43; 4{,}2]$. A titre de comparaison, sans le traitement bayésien, l'intervalle de confiance avec 95% des valeurs aurait été $[0{,}04; 3{,}96]$. La différence est significative.

Cet exemple montre comment prendre en compte à la fois des résultats de mesure et des connaissances autres pour évaluer une incertitude. Cette démarche commence à faire son apparition dans certaines normes internationales. C'est le cas, par exemple, pour la norme déjà citée sur la mesure de la radioactivité.

9. Application à une régression linéaire

Il est très aisé de trouver l'équation de la droite qui passe par deux points. Mais, quand on a trois points, ou quatre ou plus, qui ne sont pas forcément bien alignés à cause des erreurs de mesure, comment faire pour trouver la droite qui prend en compte de façon équilibrée les écarts à la droite ? Confrontés à ce que Stigler appelle un *« problème de riches »* dans son ouvrage sur l'histoire de l'incertitude avant 1900, les physiciens s'intéressant à la mécanique céleste ou à la mesure de longueur d'arcs de méridiens terrestres avaient beaucoup plus de données que de paramètres à déterminer. C'est, d'ailleurs, très souvent le cas en sciences, où l'on cherche à expliquer des phénomènes complexes avec des lois simples.

Il y a eu plusieurs tentatives de combinaison avant que Legendre, en 1805, propose un appendice sur la *« méthode des moindres quarrés »* à son mémoire sur l'orbite des comètes : *« Dans cette circonstance, qui est celle de la plupart des problèmes physiques et astronomiques, où l'on cherche à déterminer quelques élémens importans, il entre nécessairement de l'arbitraire dans la distribution des erreurs, et on ne doit pas s'attendre que toutes les hypothèses conduiront exactement aux mêmes résultats ; mais il faut sur-tout faire en sorte que les erreurs extrêmes, sans avoir égard à leurs signes, soient renfermées dans les limites les plus étroites qu'il est possible. De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre* minimum *la somme des quarrés des erreurs. Par ce moyen, il s'établit entre les erreurs une sorte d'équilibre qui empêchant les extrêmes de prévaloir, est très-propre à faire connoître l'état du système le plus proche de la vérité. »*

Point d'incertitude sur le résultat trouvé dans cette approche. Il faut attendre l'application de la théorie des probabilités. Le théorème de Bayes, retrouvé par Laplace était connu. Mais quelle distribution de probabilité choisir pour estimer l'incertitude ? Stigler décrit les différentes tentatives de Laplace. Mais c'est Gauss qui, en 1809, justifie la méthode des moindres carrés, qu'il prétend utiliser depuis 1795, en utilisant une distribution que l'on appelle gaussienne maintenant et qui avait déjà été introduite par De Moivre. Pour ce faire, il applique la version de Laplace du théorème de Bayes. Il montre aussi comment résoudre le problème, même quand il est non linéaire.

C'est le théorème central limite démontré par Laplace, qui permet à ce dernier, en 1810, de justifier l'utilisation de la distribution gaussienne et de compléter le lien entre la méthode des moindres carrés et les probabilités. Il l'appliquera ensuite à la mécanique céleste, à la géodésie et aux sciences sociales.

Ainsi, la régression linéaire découle du théorème de Bayes. On peut le vérifier aisément en langage mathématique moderne. Une théorie, un modèle ou une équation permettent de relier des paramètres p_i, comme la masse de Saturne, à des grandeurs directement observables, o_k. On peut aussi calculer directement l'incertitude théorique sur les grandeurs observables en fonction des incertitudes sur les paramètres et en déduire $P(o_k|p_i)$. Le problème qui se posait à Laplace et se pose souvent en science quand on veut déterminer des grandeurs qui ne sont pas directement mesurables, c'est exactement l'inverse : quelle est l'incertitude sur les paramètres en fonction des incertitudes expérimentales sur les grandeurs observées ? C'est, bien entendu, le théorème de Bayes qui permet de répondre.

La régression linéaire n'est qu'un cas particulier avec une loi affine qui a le mérite de pouvoir être traitée analytiquement. Supposons que, pour déterminer la pente *a* et l'ordonnée à l'origine *b* d'une droite affine, on ait *n* points de coordonnées (x_i, y_i). Supposons, de plus, qu'il n'y ait pas d'incertitude sur x_i et qu'il y en ait une sur y_i, notée u_i, et que les valeurs prises par les ordonnées y_i sont réparties selon une distribution gaussienne dont la valeur moyenne est donnée par $ax_i + b$. Ainsi,

$$P(x_i, y_i|a, b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi u_i^2}} \exp\left[-\frac{(y_i - ax_i - b)^2}{2u_i^2}\right],$$

si les grandeurs mesurées sont indépendantes. On en déduit, d'après le théorème de Bayes, que

$$P(a, b|x_i, y_i) \propto P_{prior}(a, b) \times P(x_i, y_i|a, b).$$

Le dénominateur, non écrit, correspond à la normalisation. Si l'on a aucune information sur les paramètres *a* et *b*, on va prendre un prior uniforme. Trouver les valeurs les plus probables de *a* et *b* revient à chercher les valeurs qui minimisent

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - ax_i - b)^2}{u_i^2}.$$

D'où le nom de *« moindres carrés ».* En dérivant par rapport à *a* et *b*, on peut en déduire les expressions correspondantes. Les incertitudes associées peuvent être estimées à l'aide des formules usuelles de propagation de l'incertitude.

10. Conclusion

Comme on l'a vu sur ces quelques exemples simples, la méthode initiée par Laplace est très puissante. Malheureusement, dès que l'on veut l'appliquer à des exemples un peu plus complexes, on se heurte à des calculs qui ne peuvent pas être traités analytiquement, même par Laplace. Cela a été un frein énorme.

Par ailleurs, comme on l'a vu avec le test médical, le résultat final est très sensible au choix du prior si le nombre de mesurages est faible. L'approche bayésienne a donc souffert de critiques virulentes : elle était accusée d'être trop subjective et donc non scientifique.

Les méthodes bayésiennes d'analyse des données reposent maintenant sur une axiomatique bien établie. L'avènement du calcul numérique permet aussi une application à des situations complexes. D'où l'explosion de son application à de très nombreux domaines.

11. BIBLIOGRAPHIE ET NETOGRAPHIE

BAYES Thomas, *An essay towards solving a Problem in the Doctrine of Chances*, Phil. Trans. Roy. Soc. Vol. 53 (1763) 370. Disponible en ligne : http://rstl.royalsocietypublishing.org/content/53/370

COX R.T. *Probability, Frequency and Reasonable Expectation*, Am. J. Phys. 1946, vol. 14, p. 1

BOILLEY David et LALLOUET Yoann, *Introduction aux incertitudes de mesure*, BUP Vol. 107 - Juin / Juillet / Août / Septembre 2013

GAUSS, Carl Friedrich, *Théorie du mouvement des corps célestes parcourant des sections coniques autour du soleil*, traduction par Edmond Dubois de *Theoria motus corporum cœlestium in sectionibus conicis solem ambientium*, 1809. Disponible en ligne sur http://gallica.bnf.fr/

JAYNES E.T., *Where do we stand on maximum entropy?,* http://bayes.wustl.edu/etj/articles/stand.on.entropy.pdf. Article fascinant

KAHNEMAN Daniel, *Thinking, Fast and Slow*, Penguin, 2012. Traduit en français sous le titre *Système 1 système 2, les deux vitesses de la pensée*, Paris, Flammarion, 2012.

LAPLACE Pierre Simon, *Mémoire sur la probabilité des causes par les évènements*, 1774. Disponible en ligne sur http://gallica.bnf.fr/, in Œuvres complètes, volume 8.

LAPLACE Pierre Simon, *Supplément au mémoire sur les approximations des formules qui sont fonctions de très grands nombres*, 1810. Disponible en ligne sur http://gallica.bnf.fr/, in Œuvres complètes, volume 12.

LAPLACE Pierre Simon, *Théorie analytique des probabilités,* 1ère édition en 1812, 2ième en 1814 et 3ième en 1820.

LAPLACE Pierre Simon, *Mémoire sur l'application du calcul des probabilités à la philosophie naturelle, Connaissance des temps pour l'année 1818*, 1815. Disponible en ligne sur http://gallica.bnf.fr/ in Œuvres complètes, volume 13.

LEGENDRE A. M., *Appendice sur la méthode des moindres quarrés*, in Nouvelles méthodes pour la détermination des orbites des comètes, Paris 1805. Disponible en ligne sur http://www.bibnum.education.fr.

SEGAL Jérôme, *Le Zéro et le Un : Histoire de la notion scientifique d'information au 20e siècle*, Paris, Editions Syllepse, 2003.

SHANNON Claude E., *A Mathematical Theory of Communication*, Bell System Tech. J. 1948, vol. 27, pages 379 et 623.

SIVIA D.S. *Data Analysis, a Bayesian Tutorial.* Oxford, Oxford University Press, 2006. 246 p. Ouvrage très pédagogique.

SHORE John E. et JOHNSON Rodney W., *Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy*, IEER Transactions on Information Theory 1980, vol. 26, p. 26 ; *Properties of Cross-Entropy Minimization*, IEER Transactions on Information Theory 1981, vol. 27, p. 472.

STIGLER Stephen M., *The History of Statistics, The Measurement of Uncertainty before 1900*, Cambridge, Massachusetts, The Belknap press of Harvard University press 1986. Excellent livre qui a inspiré la partie historique de cet article.